Credit EDA Assignment

By Vivek Pal

Understanding:

- If the applicant is likely to repay the loan, then not approving the loan is results in a loss of the bank or to the company.
- If the applicant is not likely to repay the loan amount, they will be defaulters, than approving loan to them will be a finencial loss of the bank or Company.

This analysis help to identify patterns which will show if a client has difficulty paying their installments which may be used to taking actions such as denying the loan, reducing the amount of the loan, reducing the amount of the loan, lending at a higher interest rate, etc… This will be ensure that the consumers are capables to repay the amount are not should be rejected.
Identification of such applicant's using EDA is the aim of this Analysis.

Data Understanding:

Application_Data.csv (data)	Previous_Application_Data_csv (prvs_data)
- Number of Columns - 122 - Number of Raws - 3,07,511 - Data Types - Integers, Float, & Strings - Descriptive view od data file: There were anomolies like negative numbers, Null values, Days, and Years were not in proper format. ❖ Float64: 64 ❖ Int64: 41 ❖ Object 16	- Number of Columns - 37 - Number of Raws - 16,70,214 - Data Types - Integers, Float, & Strings - Descriptive view od data file: There were anomolies like negative numbers, Null values, Days, and Years were not in proper format. ❖ Float64 - 15 ❖ Int64 - 06 ❖ Object - 16

Data Cleaning & Manipulations for Application Data:

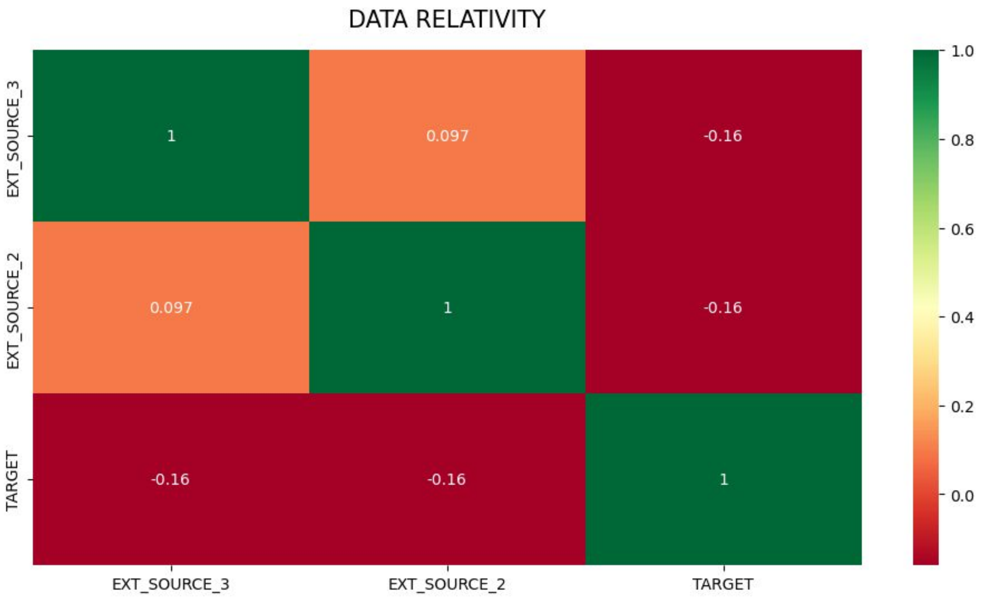
- Rectify the null values
- Filtering unwanted data columns
- Filling the missing values
- Sorting the data
- Fixing the datatype

To remove unwanted or irrelevant columns,

- First, have calculated null values “nulls(data)”
- Then, Calculated the values in term of %
- Found that there were above 40+ columns which consists more than 50% null values
- By comparing the columns with given csv’s file, Removed the irrelevant columns
- Similarily, after removing the 50% data there were 10 columns which were Null more than 15% null values

Data Cleaning & Manipulations for Application Data:

- After double check those 15% null values, There were outsourced data columns which are provided by externally.
- Source Columns: EXT_SOURCE_2 & EXT_SOURCE_3
- What is the relation between these 1 values> As per the column description datafile, These are normalized values from external data.



Data Cleaning & Manipulations for Application Data:

- By above mentioned correlation heatmap, we found that there were no relation and not much contribution.
- These data doesn’t cause anything
- So, on this base i have removed the EXT_Source_2 & EXT_Source_3 Columns

After removing all these columns, we left with 116 Columns

These 116 columns includes 28 flag columns

- In which there are Emails, Phone No, Car, Work, and Other important data were stored.
- To analyze the Flag data, i have combined all the flag columns in one variable “col_flag”
- Includes, “Target” variables, which has explains
- For analysis we need to find the Payers & Defaulters, for that have changed data from 1’s 0’s to “Defaulters” & “Payers”

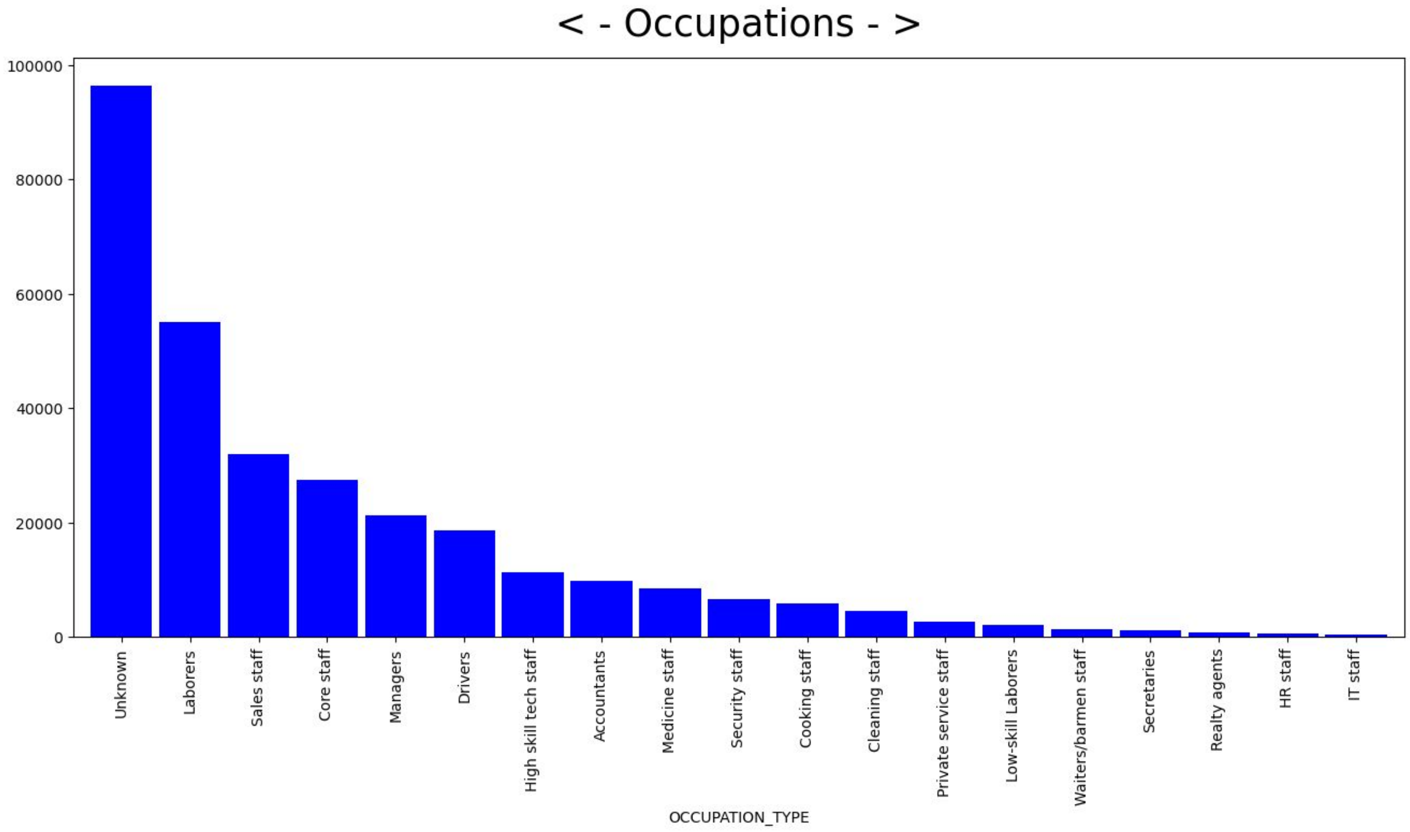
Analyzing Flag columns & Target Columns:

1. By observing the Graphs:
2. Defaulter
3. A. (FLAB_OWN_REALTY)
   B. FLAG_MOBIL
   C. FLAG_EMP_PHONE
   D. FLAG_CONT_MOBILE
   E. FLAG_DOCUMENT_3
4. These columns make relativity, we can include these below:

a. FLAG_DOCUMENT_3
b. FLAG_OWN_REALITY
c. FLAG_MOBIL

5. We can remove all other Flag columns.

Imputing Values



- In the 10 Columns, there was a column "OCCUPATION_TYPE" which describes the "USER OCCUPATION" was having 31% of null values.
- I have used "Unknown" variables to fill those 31% null values
- First highest percentage is "Unknown"
- Second highest % is labours

Standardizing the Values:

- Very high value data columns:

a. AMT_INCOME_TOTAL,
b. AMT_CREDIT,
c. AMT_GOODS_PRICE

- Converting these numerical columns in categorical columns for better understanding

Negative values data columns:

a. DAYS_BIRTH
b. DAYS_EMPLOYED
c. DAYES_REGISTRATION
d. DAYS_ID_PUBLISH
e. DAYS_LAST_PHONE_CHANGE

Need to make it correct those values convert DAYS_BIRTH to AGE in years, DAYS_EMPLOYED to YEARS_EMPLOYED.

Standardizing the Values:

Standardizing AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE Column:

- Its has pricing from 0 to lakhs, so, mad category and divide the pricing
- “Income Range” range from 0 to 10 lakhs

  Bins - [0,1,2,3,4,5,6,7,8,9,10,11]

  Slot - [‘0-1L’, ’1L-2L’, ’2L-3L’, ’3L-4L’, ‘4L-5L’, ‘5L-6L’, ‘6L-7L’, ‘7L-8L’, 8L-9L’, ‘9L-10L’, ‘10L-Above’]

- Made “Credit Range” range from 0 to 10 Lakhs

  Bins - [0,1,2,3,4,5,6,7,8,9,10,100]

  Slot - [‘0-1L’, ’1L-2L’, ’2L-3L’, ’3L-4L’, ‘4L-5L’, ‘5L-6L’, ‘6L-7L’, ‘7L-8L’, 8L-9L’, ‘9L-10L’, ‘10L-Above’]

- Made “Price of Goods” range from 0 to 10 Lakhs

  Bins - [0,1,2,3,4,5,6,7,8,9,10,100]

  Slot - [‘0-1L’, ’1L-2L’, ’2L-3L’, ’3L-4L’, ‘4L-5L’, ‘5L-6L’, ‘6L-7L’, ‘7L-8L’, 8L-9L’, ‘9L-10L’, ‘10L-Above’]

Standardizing the Values:

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
count	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000
mean	-16036.995067	63815.045904	-4986.120328	-2994.202373	-962.858788
std	4363.988632	141275.766519	3522.886321	1509.450419	826.807143
min	-25229.000000	-17912.000000	-24672.000000	-7197.000000	-4292.000000
25%	-19682.000000	-2760.000000	-7479.500000	-4299.000000	-1570.000000
50%	-15750.000000	-1213.000000	-4504.000000	-3254.000000	-757.000000
75%	-12413.000000	-289.000000	-2010.000000	-1720.000000	-274.000000
max	-7489.000000	365243.000000	0.000000	0.000000	0.000000

Before + ve Values

- As mentioned above -
- Negative values Data Columns:

  DAYS_BIRTH
  DAYS_EMPLOYED
  DAYS_REGISTRATION
  DAYS_ID_PUBLISH
  DAYS_LAST_PHONE_CHANGE

Standardizing the Values:

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
count	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000
mean	16036.995067	67724.742149	4986.120328	2994.202373	962.858788
std	4363.988632	139443.751806	3522.886321	1509.450419	826.807143
min	7489.000000	0.000000	0.000000	0.000000	0.000000
25%	12413.000000	933.000000	2010.000000	1720.000000	274.000000
50%	15750.000000	2219.000000	4504.000000	3254.000000	757.000000
75%	19682.000000	5707.000000	7479.500000	4299.000000	1570.000000
max	25229.000000	365243.000000	24672.000000	7197.000000	4292.000000

- As mentioned above -
- Negative values Data Columns:

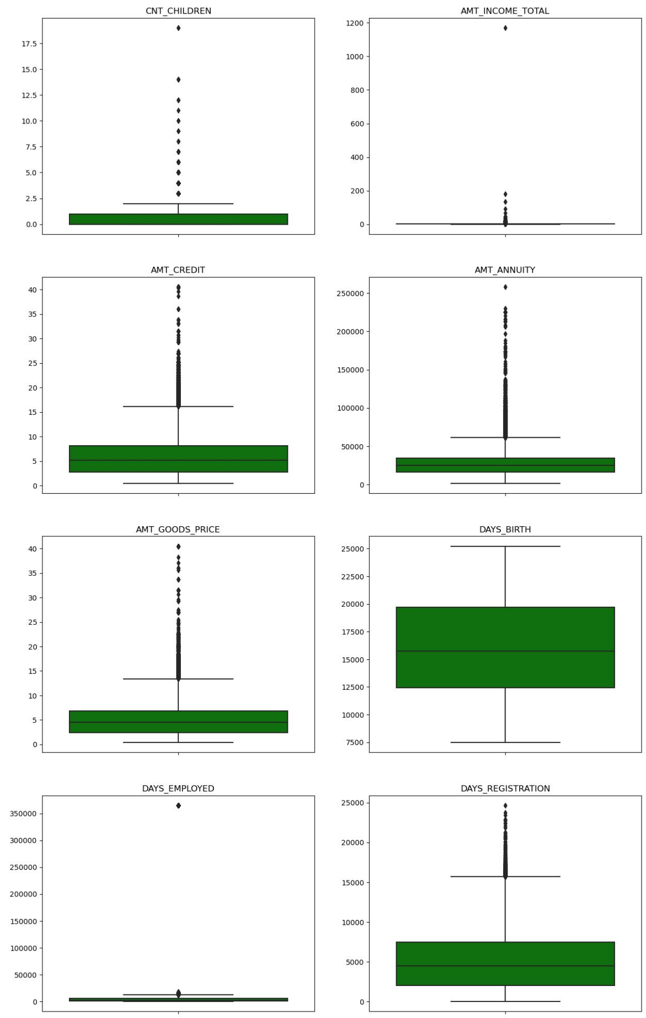
  DAYS_BIRTH
  DAYS_EMPLOYED
  DAYS_REGISTRATION
  DAYS_ID_PUBLISH
  DAYS_LAST_PHONE_CHANGE

After + ve Values

Standardizing the Values:

Find the outliers

- Max Outliers: AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, CNT_CHILDREN
- Min Outliers: AMT_INCOME_TOTAL
- No Outliers: DAYS_BIRTH



Summary on Datasets: Application_Data.csv

States that: Application_Data.csv

There are 3,07,511 Raws and 97 Columns

Types of Datatypes available

- Integers
- Float Values
- Strings

Found the Null values, filled them with “Unknown” variable

Removed unwanted columns & other columns

Worked on the negative values and converted them into positive values in some of columns

I have converted values in proper format

Now, file is neat & clean for further process.

Summary on Datasets: Previous_Application_Data.csv

States that: Application_Data.csv

There are 1670214 Raws and 37 Columns

Types of Datatypes available

- Integers
- Float Values
- Strings

Found the Null values, filled them with “Unknown” variable

Removed unwanted columns & other columns

Worked on the negative values and converted them into positive values in some of columns

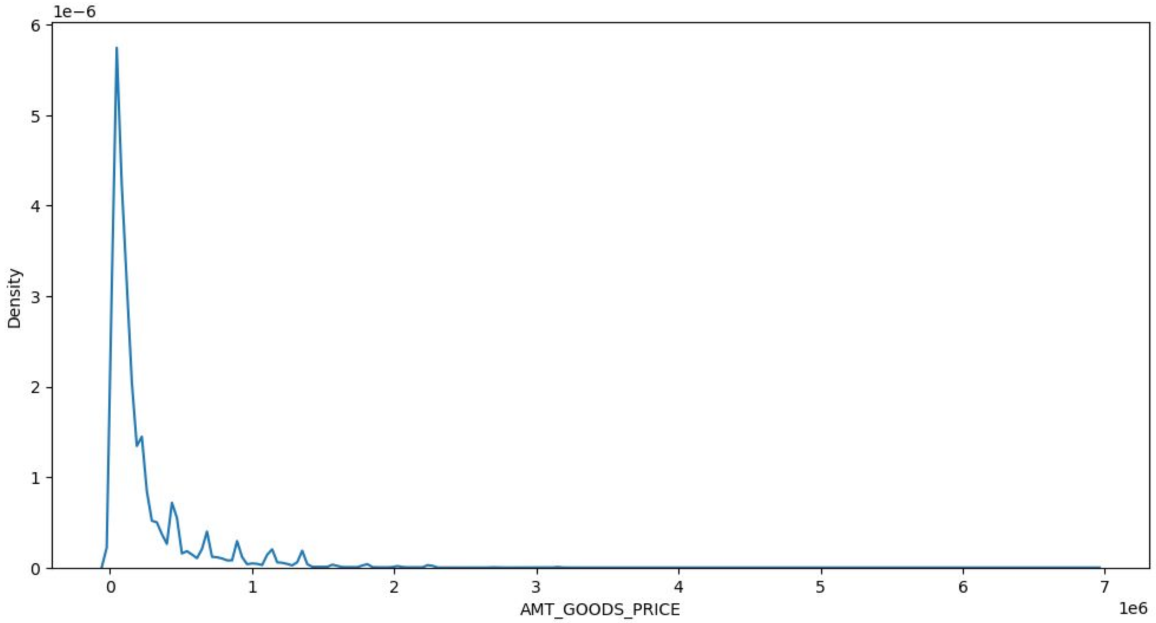
I have converted values in proper format

Now, file is neat & clean for further process.

Data Set Analyzing Using Graphical Representations

Analyzing the Data using Kdeplot:

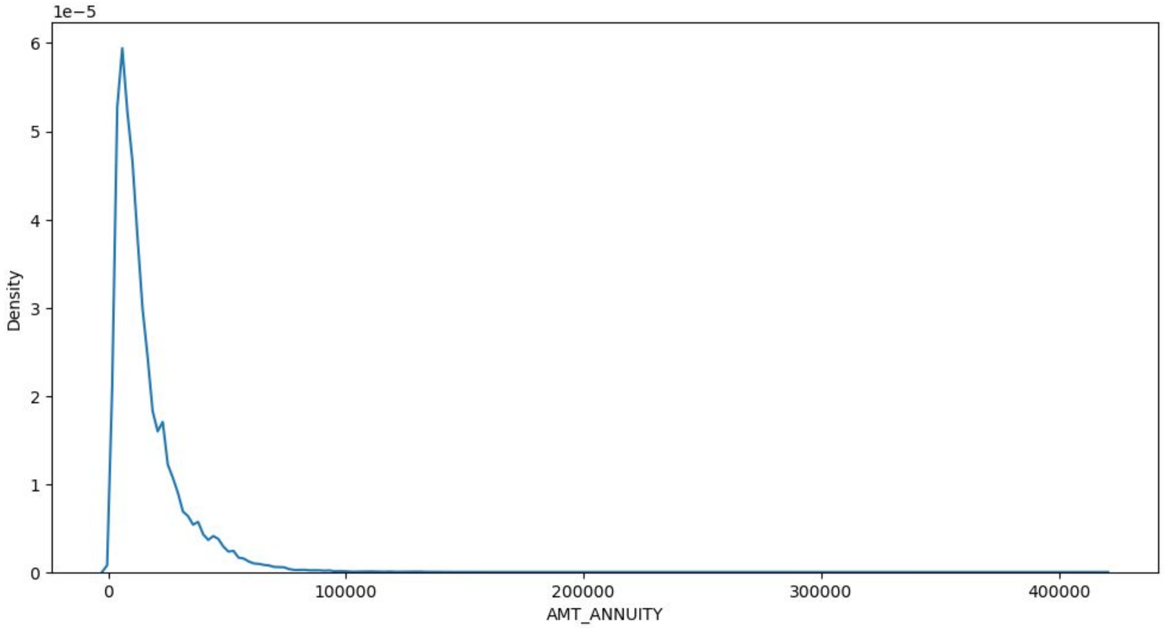
1. Plotting kde for "AMT_GOODS_PRICE" to understand the distribution.
2. There were several peaks along the distribution, Let's impute using the mode, mean, and median, and see if the distribution is still about the same.



Data Set Analyzing Using Graphical Representations

Analyzing the Data using Kdeplot:

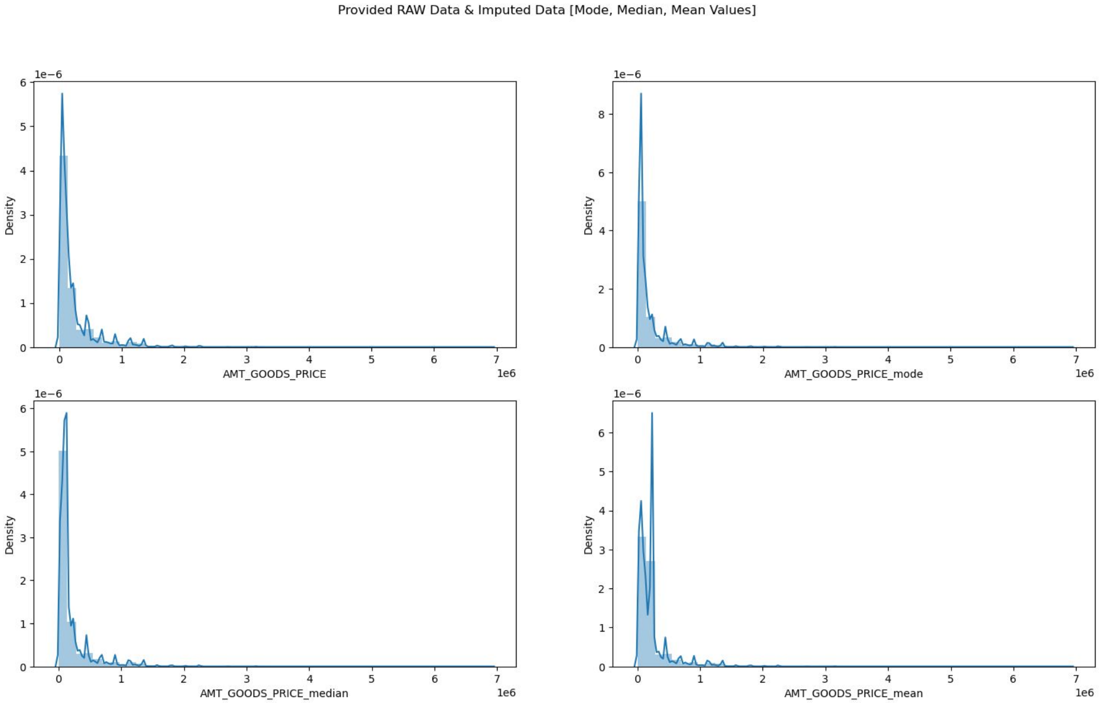
1. Plotting kde for “AMT_ANNUITY” to understand the distribution.
2. There were single peaks along the distribution, Let’s impute using the mode, mean, and median, and see if the distribution is still about the same.



Data Set Analyzing Using Graphical Representations

Analyzing the Data using Kdeplot:

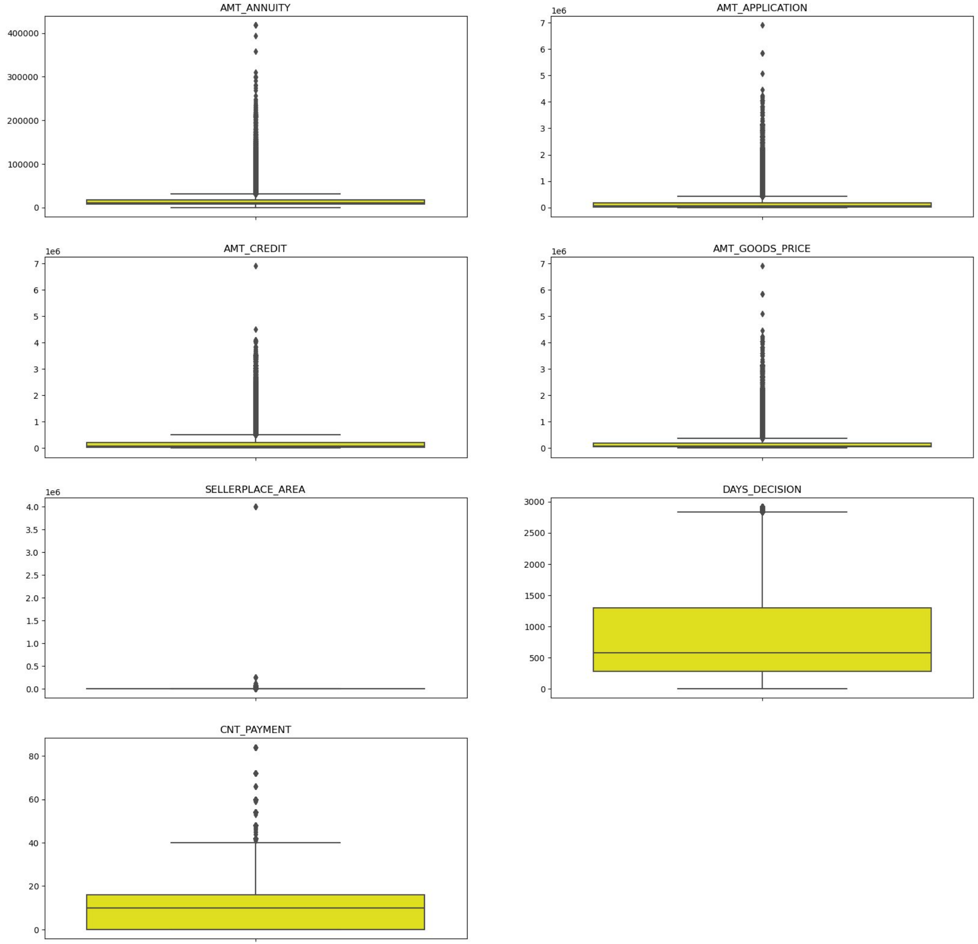
The Original distribution is closer with the distribution of data imputed with mode in this case, thus will impute mode for missing value.



Finding outliers In:

[‘AMT_ANNUITY’, ‘AMT_APPLICATION’, ‘AMT_CREDIT’, AMT_GOOODS_PRICES’, ‘SELLERPLACE_AREA’, ‘DAYS_DECISION’, ‘CNT_PAYMENT’]

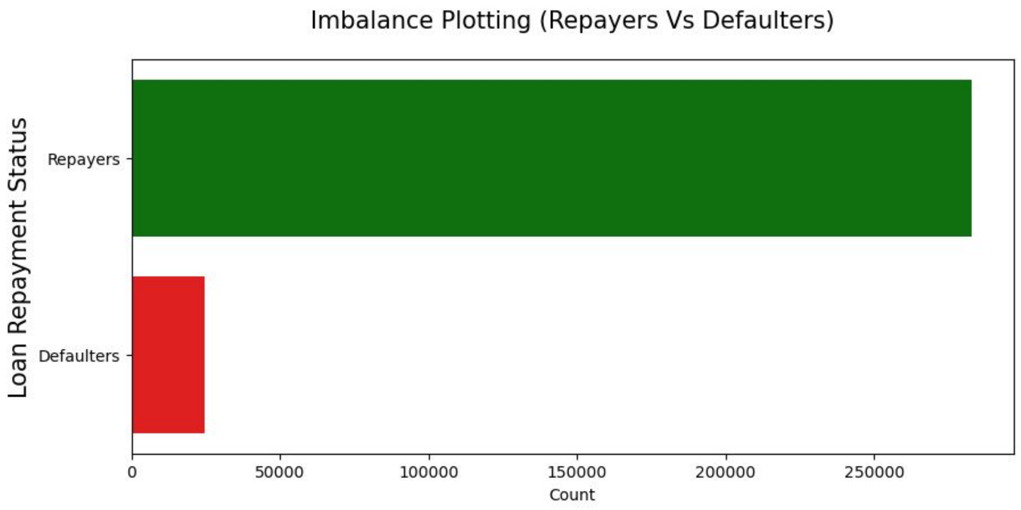
- Summary - It can be seen that previous application data
- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA, consists max number of outliers.
- CNT_PAYMENT has little number of outliers indicating that these previous application decision.



Data set analyzing using Graphical Representation

Repayers & Defaulters -

- Repayers % is 91.93%
- Defaulter % is 8.07%
- Imbalance ratio with respect to Repayers & Defaulters is given: 11.39/1



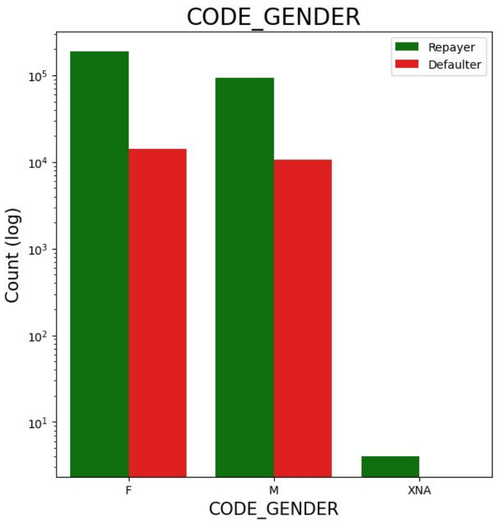
Data set analyzing using Graphical Representation

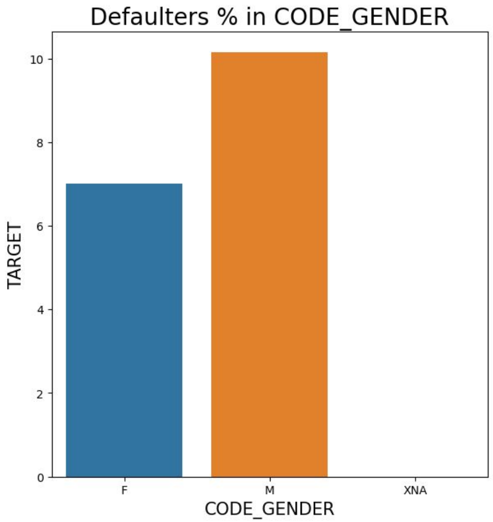
Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

Gender wise Analysis

Based on the percentage of default credits, males have a higher chance to not returning their loans, comparing to women.





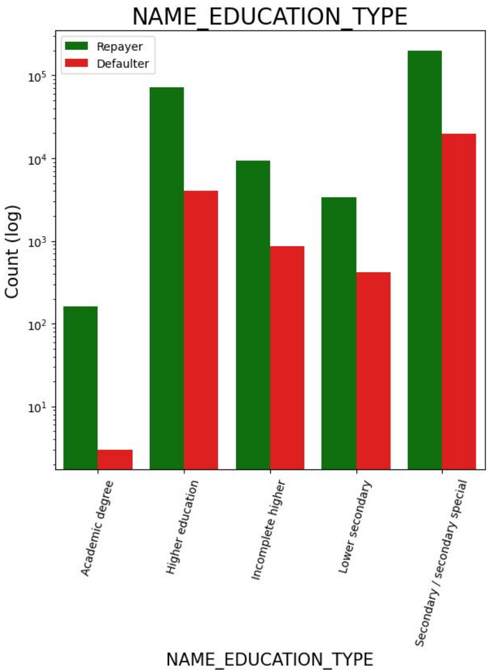
Data set analyzing using Graphical Representation

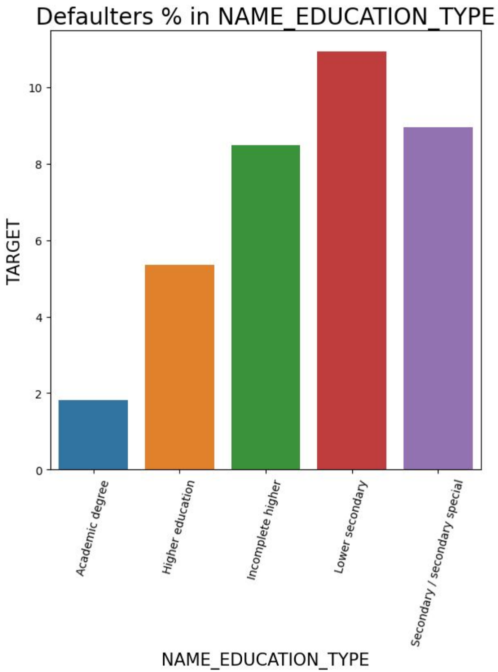
Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

Education wise Analysis

Majority of clients have Secondary/secondary special education, followed by clients with Higher education. Very few clients have an academic degree Lower secondary category have highest rate of defaulter. People with Academic degree are least likely to default.





Data set analyzing using Graphical Representation

Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

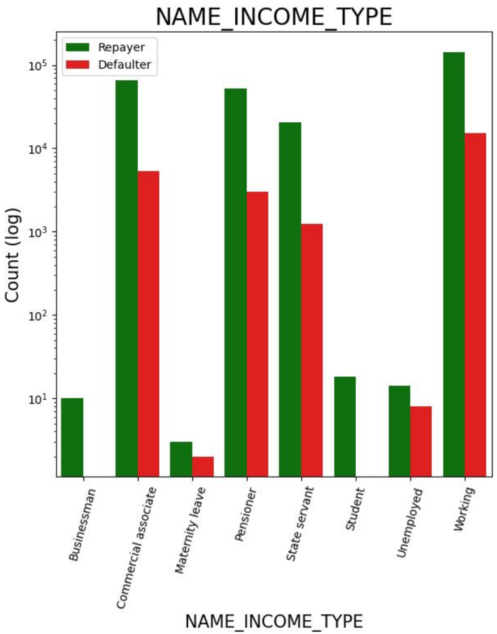
Income wise Analysis

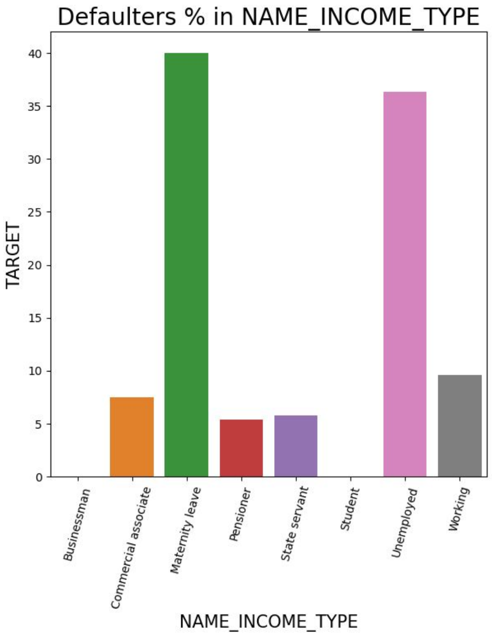
Most of applicants for loans income type is Working, followed by Commercial associate, Pensioner and State servant.

The applicants who are on Maternity leave have defaulting percentage of 40% which is the highest, followed by Unemployed (37%).

The rest under average around 10% defaultees.

Student and Businessmen though less in numbers, do not have default record. Safest two categories for providing loan.





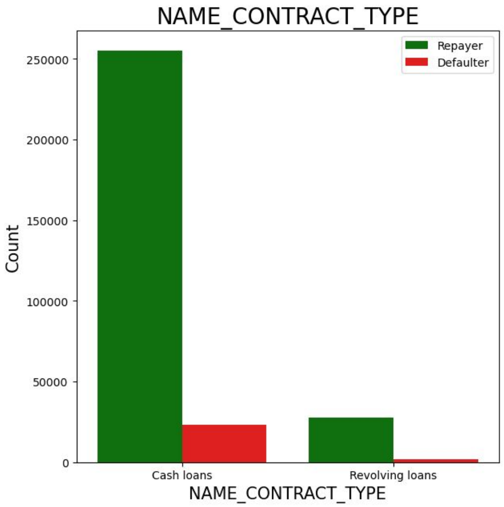
Data set analyzing using Graphical Representation

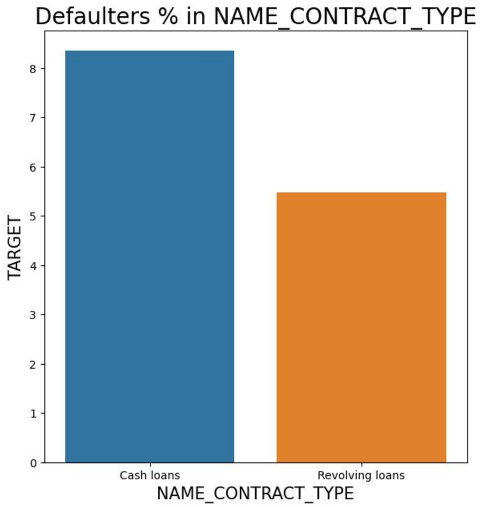
Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

Contract wise Analysis

Contract type: Revolving loans are just a small fraction (10%) from the total number of loans Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters





Data set analyzing using Graphical Representation

Analyzing, Univariate, Bivariate, Multivariate

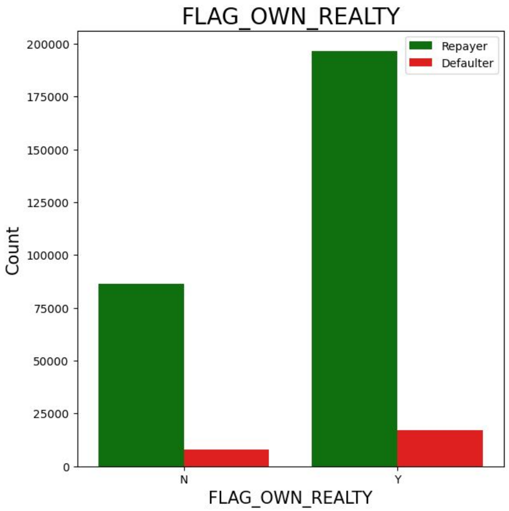
Categorical Univariate Variables Analysis -

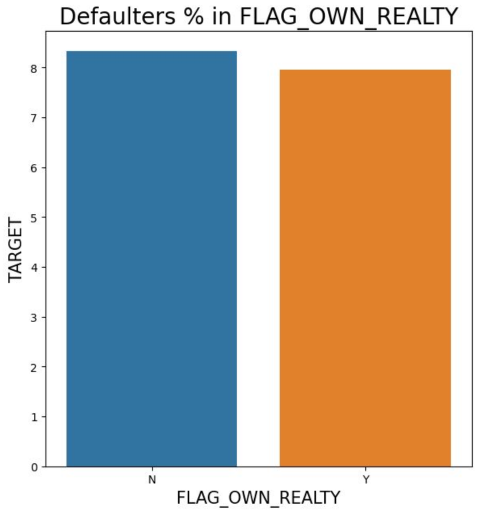
Real Estate wise Analysis

The clients who own real estate are more than double of the ones that don't own.

The defaulting rate of both categories are around the same (~8%).

Thus we can infer that there is no correlation between owning a reality and defaulting the loan.





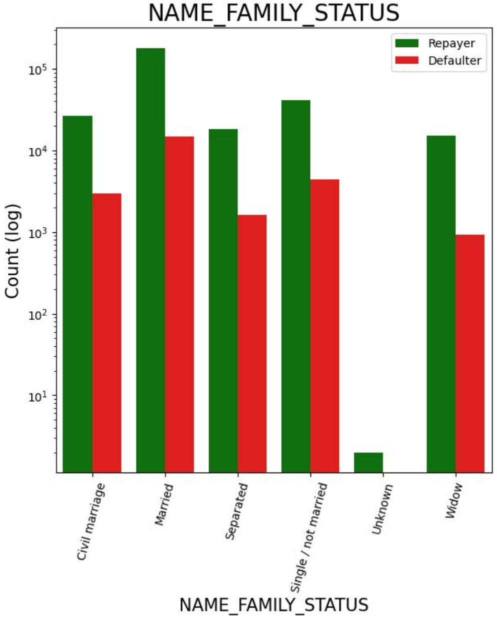
Data set analyzing using Graphical Representation

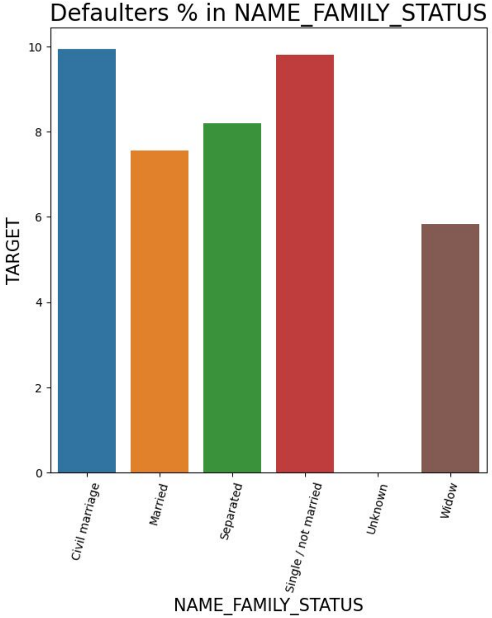
Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

Occupation wise Analysis

Most of the people who have taken loan are married, followed by Single/not married and civil marriage. In Percentage of defaulters Civil marriage has the highest percent around and widow has the lowest.





Data set analyzing using Graphical Representation

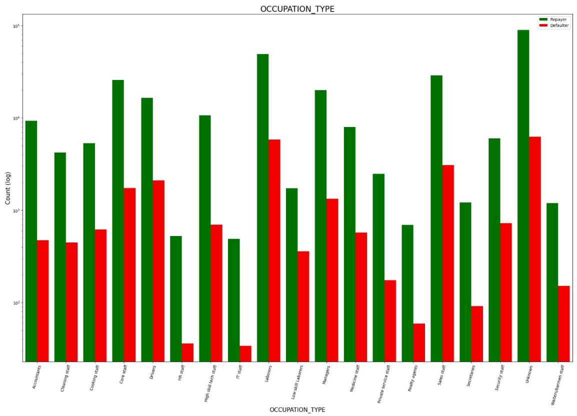
Analyzing, Univariate, Bivariate, Multivariate

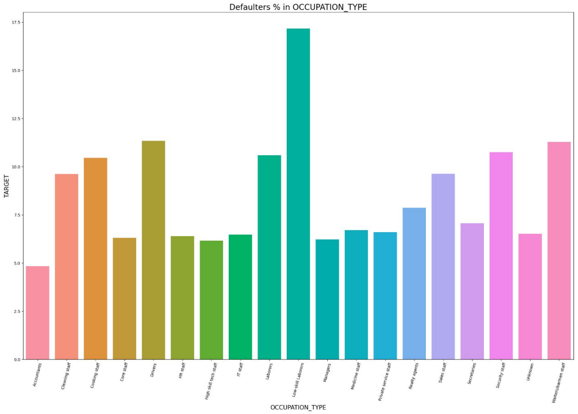
Categorical Univariate Variables Analysis -

Occupation Analysis

Category with highest percent of defautess are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.

IT staff are less likely to apply for Loan.



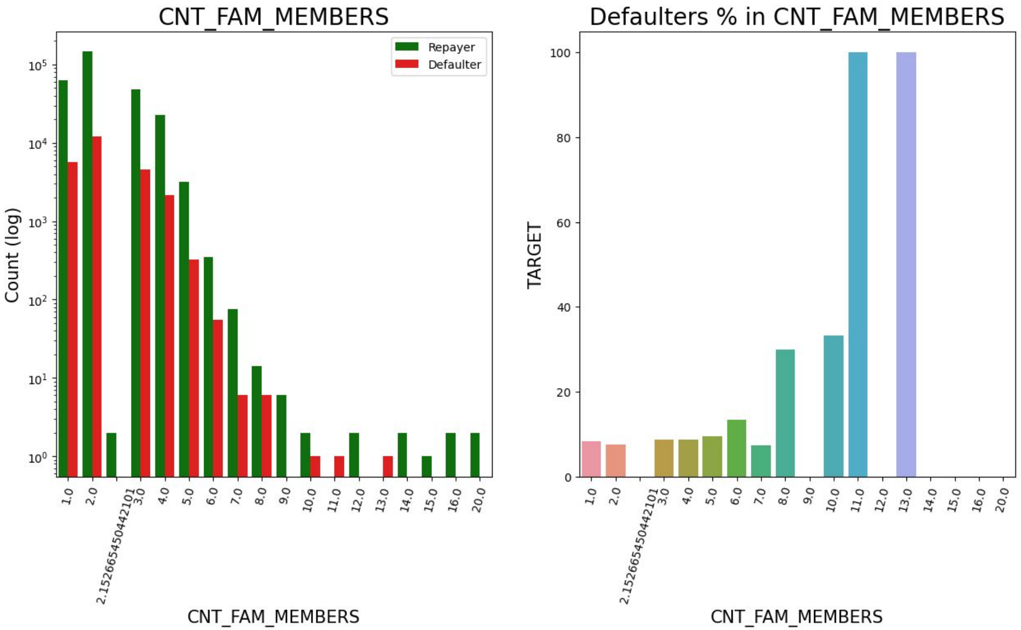


Data set analyzing using Graphical Representation

Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

Number of Families Analysis



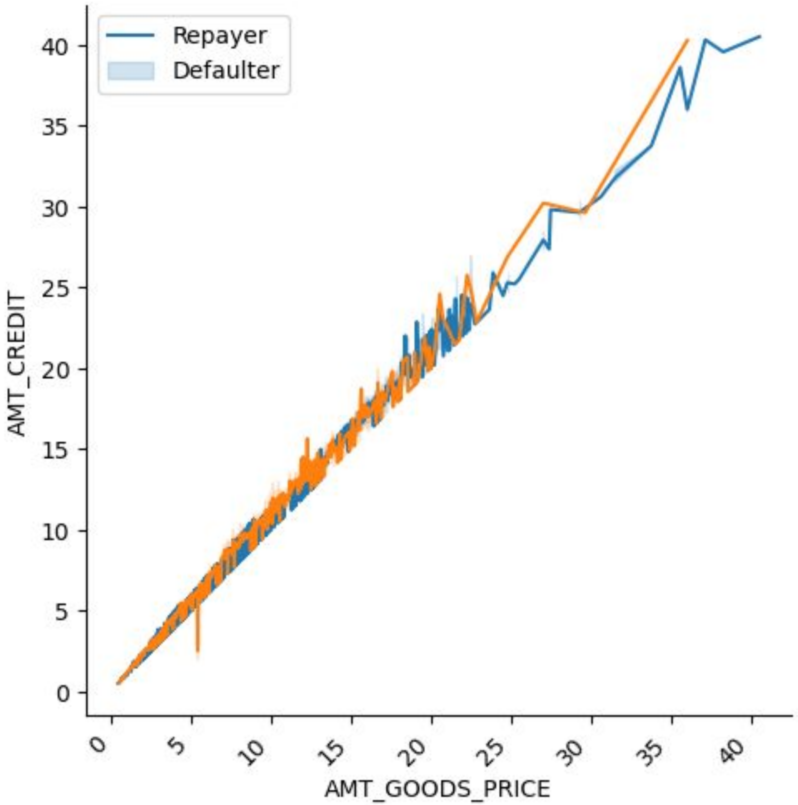
Data set analyzing using Graphical Representation

Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

Numerical Univariate Analysis

When the credit amount goes beyond 30 Lakhs, there is an increase in defaulters.



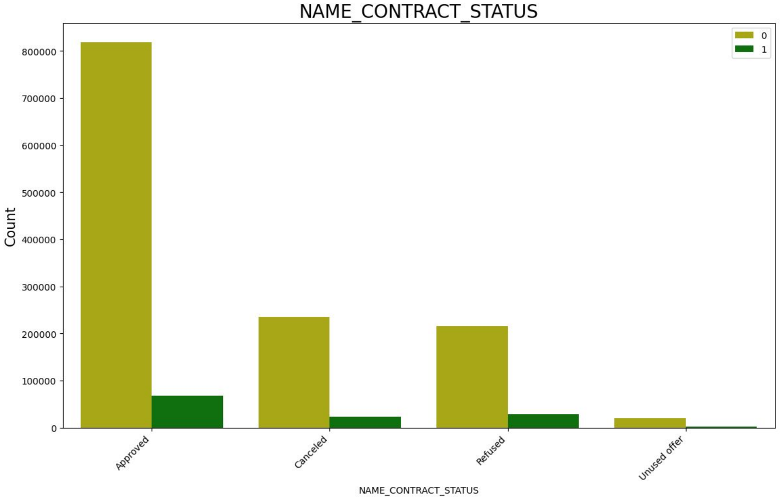
Data set analyzing using Graphical Representation

Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

Numerical Univariate Analysis

90% of the previously cancelled client have actually repayed the loan. Revising the interest rates would increase business opportunity for these clients88% of the clients who have been previously refused a loan has payed back the loan in current case.Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.



NAME_CONTRACT_STATUS	TARGET	Counts	Percentage
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

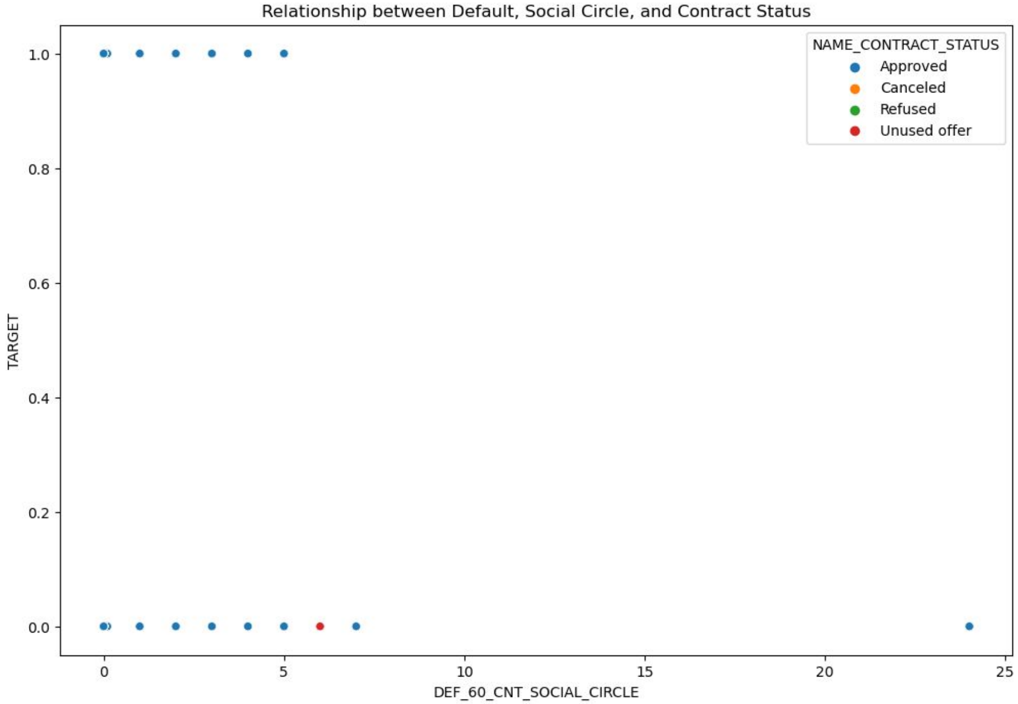
By Vivek Pal

Data set analyzing using Graphical Representation

Analyzing, Univariate, Bivariate, Multivariate

Categorical Univariate Variables Analysis -

Clients who have an average of 0.13 or higher DEF_60_CNT_SOCIAL_CIRCLE score tend to default more, and thus analyzing the client's social circle could help in the disbursement of the loan.



Thank You!

By Vivek Pal